DOCUMENT RESUME

ED 441 841 TM 030 876

AUTHOR Yang, Wen-Ling; Chen, Wen-Hung
TITLE Estimating Cutpoints for the Achievement-Levels Setting Process for the National Assessment of Educational Progress.
SPONS AGENCY American Institutes for Research (CRESS), Kensington, MD.; National Assessment Governing Board, Washington, DC.
PUB DATE 2000-04-00
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
CONTRACT ZA97001001; ZA93003001; RN91226001
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Academic Standards; Constructed Response; *Cutting Scores; Elementary Secondary Education; *Estimation (Mathematics); *Item Response Theory; Multiple Choice Tests; National Competency Tests; *Scoring
IDENTIFIERS *National Assessment of Educational Progress; *Standard Setting

ABSTRACT

This paper provides a historical review of the changes and improvements made in estimating numerical cutpoints for the National Assessment of Educational Progress (NAEP). While reviewing the various methodologies used for collecting judgment data, the paper discusses: (1) the incorporation of Item Response Theory for setting standards; (2) methodologies for converting judgment data into achievement level cutpoints; (3) combining cutpoints estimated for dichotomously scored multiple-choice items and polytomously scored constructed response items; and (4) deriving group cutpoints from the estimates of individual cutpoints for items and judges. In addition to the eight formal achievement level setting (ALS) processes that have been used, pilot studies and field trials have evaluated other approaches before the formal standard-setting process. In general, methodologies can be categorized into item-by-term rating (analytical) and holistic approaches. This review illustrates the importance and significance of the application of Item Response Theory in setting NAEP achievement levels. (Contains 1 table and 25 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 841

Estimating Cutpoints for the Achievement-Levels Setting Process For the National Assessment of Educational Progress

By

Wen-Ling Yang, Educational Testing Service
Wen-Hung Chen, American Institutes for Research

April, 2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030876

Paper prepared for the presentation at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

The research reported in this paper was supported by contracts ZA97001001, ZA93003001, and RN91226001 between ACT and the National Assessment Governing Board.

BEST COPY AVAILABLE

Estimating Cutpoints for the Achievement-Levels Setting Process For the National Assessment of Educational Progress

Introduction

The policy of the National Assessment Government Board (NAGB) on establishing achievement levels for the NAEP required that three cutpoints be set on the NAEP score scale for each of the three grades tested (4, 8, and 12) and for each NAEP test title (NAGB, 1990). These cutpoints should correspond to the definitions for the achievement levels of Basic, Proficient and Advanced, approved by NAGB. The establishment of the achievement-levels cutpoints should improve the reporting of the NAEP assessment results by providing data on the proportion of students achieving each achievement level. The task requires mapping the estimated three achievement levels onto the NAEP score scale for each grade. Hence, methodologies and technical procedures were developed and used for estimating achievement level cutpoints. With considerations of the NAEP context and the development of the method used for collecting judgement data, the techniques for estimating cutpoints from empirical data evolved over years.

To address the evolution of the Achievement-Levels Setting (ALS) for the National Assessment of Educational Progress (NAEP), on a more technical side, this paper provides a historical review of the changes and improvements in estimating numerical cutpoints for the NAEP. While reviewing the methodologies used for collecting judgement data, this paper discusses these issues in retrospect:

- The incorporation of Item Response Theory (IRT) for setting standards.
- Methodologies for converting judgement data into achievement-level cutpoints.
- Combining cutpoints estimated for dichotomously scored multiple-choice items and polytomously scored constructed-response items.
- Deriving group cutpoints from the estimates of individual cutpoints for items and judges.

Overview of Methodologies for the NAEP ALS Process

ACT has been awarded with several contracts from NAGB for developing achievement levels for the NAEP tests since 1992. The NAEP ALS project, however, began in 1990 and achievement levels were set for the NAEP Mathematics of the same year. The ALS process for the 1990 Mathematics NAEP was conducted on a relatively small scale, and a modified-Angoff method was used for collecting judges' item rating data (Hambleton & Bourque, 1991).

In addition to the eight formal ALS processes, in the past nine years ACT also proposed and implemented pilot studies and/or field trials before the formal standard setting processes. These studies provided opportunities for experimenting and evaluating innovative methods and techniques for setting standards. After each of the ALS processes, analyses and validation studies were conducted to gather validity evidence of the ALS cutpoints. Table 1 summarizes the methodologies developed and implemented for each of the NAEP ALS-related studies conducted by ACT.

Standard Setting Methodologies Implemented for the ACT NAEP Achievement-Levles Setting (ALS) Processes

NAEP Test	Study	Item Type	Methodology
1992 Mathematics	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Paper Selection
1992 Reading	Pilot	Dichotomous	Modified-Angoff Method
		Polytomous	Paper Selection
	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Paper Selection
1992 Writing	ALS	Polytomous	Paper Selection
1994 Geography	Pilot	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
			Estimation of Score-Point Percentages
			The Hybrid Method
	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	Validation (1995)	All	Booklet Classification Method
			Similarities Classification Method*
1994 U.S. History	Pilot	Dichotomous	Modified-Angoff Method
		Polytomous	Estimation of Score-Point Percentages
			Modified Percentage Estimate
			The Hybrid Method
	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	Validation (1995)	All	Booklet Classification Method
			Similarities Classification Method*
1996 Science	Pilot #1	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	Pilot #2	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	ALS Reconvening	All	Item Mapping (Grade 8 Only)
	Validation (1997)	All	Booklet Classification Method
1998 Writing	Field Trial #1	Polytomous	Mean Scores Estimation
			ISSE Method
	Field Trial #2		Booklet Classification Method
			Reckase Method
	Pilot		Mean Scores Estimation
	ALS		Mean Scores Estimation
1998 Civics	Field Trial #1	All	ISSE Method
	Field Trial #2	Dichotomous	Modified Angoff (Round 1 & Round 2) with Item Maps (Round 3)
		Polytomous	Modified Angoff (Round 1 & Round 2) with Item Maps (Round 3)
		All	Reckase Method
	Pilot	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	ALS	Dichotomous	Modified-Angoff Method
		Polytomous	Mean Scores Estimation
	Validation (1998)	All	Booklet Classification Method
			Similarities Classification Method*

* The Similarities Classification Method was developed to collect indirect validity evidence for the numerically defined NAEP Achievement Levels. Special reports have been written for its application in the NAEP ALS validation studies. Technically, this method is not designed or used for setting achievement levels. Therefore, it will not be discussed in the present paper.

In general, the merit of the methodologies experimented for the NAEP ALS processes were evaluated from these perspectives:

- Theoretical and statistical appeals of the method.
- Applicability of the methodology in the context of the NAEP.
- Reasonableness of the resulting achievement-level cutpoints and the inter-judge agreement.
- Corresponding percentage of students at or above each achievement level.
- Intra-judge rating consistency across rating items within and across rounds of ratings, if the method is based on item-by-item rating.
- Evaluation outcomes from the standard setting panelists regarding the ALS process and achievement-level outcomes.

ACT has implemented a variety of methodologies for setting achievement levels for the NAEP, as shown in Table 1. These methodologies can be categorized into two methodological groups: item-by-item rating (analytic) approach and holistic approach. To provide methodological background for the computation procedures for the ALS cutpoints, this paper briefly summarizes the development of the standard setting methodologies in the following section. Basically, the rating methods used for the 1992 ALS were the paper selection and the modified Angoff methods. Starting in 1994, the methods used were the mean scores estimation and the modified Angoff methods. Item Response Theory (IRT) is an important and integral part of the ALS process. In terms of computing the numerical achievement-level cutpoint, the significant efforts include the use of the test characteristic curve (TCC), the information weighting technique for combining the cutpoints for different types of items or content domains, and the MLE approach.

Item-by-Item Rating (Analytic) Approach

Many of the ALS methods were based on a panelist's judgement about student performance on individual items. Therefore, the cutpoint estimation procedure for the item-by-item rating approach usually starts with estimating the score (or probability of a correct response) of a borderline performance. Then, the estimated scores (or probabilities) across items were combined through some mechanism such as IRT to get an overall estimate for the cutpoint for a panelist or a group of panelists.

Modified-Angoff Method

The modified-Angoff method (Angoff, 1971; Jaeger, 1989) was chosen for setting achievement levels for the Mathematics NAEP when the NAEP ALS first started in 1990. The modified Angoff method was further "modified" in the sense that three achievement levels in their numerical representations were to be mapped on an achievement scale, which would be further transformed to the NAEP score scale for reporting. The method was chosen and applied for these reasons (Hambleton & Bourque, 1991):

- The literature on standard setting indicated that the modified-Angoff method was generally superior to the other competing procedures.
- The procedure was fairly straightforward and the interpretation of its results was not difficult.
- The procedure did not require the administration of items to a trial population.

The modified-Angoff method calls for an item-by-item rating process, where standard-setting panelists will provide probabilities of correct response for a minimally acceptable student at each of the achievement levels. The method involves multiple rounds of panelist ratings, with discussion and feedback between rounds. The purpose is to generate more reliable ratings, within and between panelists, to obtain estimates of cutpoints with smaller estimation error than the one-round rating process.

Since 1992, the modified-Angoff method and its statistical estimation techniques used for deriving cutpoints from panelist ratings have been further modified over years. The modifications were to accommodate the IRT calibration practice for the NAEP test items and ability score estimates. Nevertheless, the rationale and principles of the item-by-item ratings remained. In short, the cutpoint estimation technique evolved from projecting the average panelist ratings onto a latent trait (θ) scale via the test characteristic curve (TCC) to obtaining the θ via maximum likelihood estimation based on the item ratings. These techniques are further discussed in this paper.

Mean Scores Estimation Method

For setting standards for the polytomously scored NAEP items, the modified-Angoff rating method was extended to account for the probabilities associated with all of the possible score points for an item. Exploratory studies were conducted to evaluate the applicability of these extensions, with theoretical considerations (Luecht, 1993a). In general, these extensions of the modified-Angoff method were not appealing because their rating tasks were very complex for the standard setting panel. Also, there was a lacking of methodological investigation in practice. In the early years of the ALS process, ACT had used a holistic approach--the paper selection method (discussed later in this paper)--to set standards for the NAEP performance items. However, this method also suffered from operational limitations and had theoretical drawbacks.

Consistent with the IRT calibration in the NAEP context and in consideration of the reasonable level of cognitive demand for rating items, ACT proposed and implemented the Mean Scores Estimation method (ACT, 1994). The method replaced the paper selection method for setting achievement levels for the polytomously scored items. The Mean Scores Estimation method requires panelists to estimate the mean score for students at the lower borderline of each achievement level for each item. Compared to the extended modified-Angoff methods, the Mean Scores Estimation reduces the workload for panelists by lowering the degree of complexity for item rating. Relative to the Paper Selection method, the implementation of the Mean Scores Estimation method has less operational difficulties.

Estimation of Score-Point Percentages & the Modified Percentage Estimate

The estimation of score-point percentages method (ACT, 1994) has panelists estimate the percentage of borderline students at each achievement level, who would be scored at each point on the score scale for the polytomously scored item. The method of Modified Percentage Estimate (ACT, 1994) was a variation of the Estimation of Score-Point Percentages. It requires the estimation of the percentage of borderline students, at each achievement level, who would be scored at score point of two or higher.

The above two methods were applied for the polytomously scored items. The obtained percentage/probability estimates were mapped through the expected score function for each item to derive an estimate of a judge's standard (θ) for each score point of an item. Then, the θ estimates were used to estimate the expected scores for each item, which were then averaged across items and panelists. The overall average expected scores were mapped via the test characteristic curve to estimate the achievement level cutpoints (Luecht, 1993b).

Although the Estimation of Score-Point Percentages and its modified version were experimented in the pilot studies for the 1994 NAEP Geography and History, neither method was used for the formal ALS process. The results of the pilot studies indicated that the applicability of these two methods needed to be further investigated.

The Hybrid method

The Hybrid method (ACT, 1994) combines the paper selection process and the Mean Scores Estimation method. The Paper Selection process was used for the first round of ratings to familiarize panelists with actual student responses to the constructed response items. For the second and the third rounds of ratings, the less time-consuming Mean Scores Estimation method was used.

The Hybrid method was tried out in the pilot studies for the 1994 NAEP Geography and History. Outcomes from the pilot studies indicated that the paper selection part of the process was very time-consuming. In addition, the resulting achievement levels might not be different from the results produced by the Mean Scores Estimation method. Therefore, the Hybrid method was not used for the formal ALS process.

The ISSE Method

The Item Score String Estimation (ISSE) method was proposed as an alternative rating method for the modified-Angoff procedure. One of the criticisms about the modified-Angoff method is that it fails to produce valid numerical cutpoints because panelists might not be able to perform the required task of estimating probabilities with reasonable accuracy. The rating task required by the ISSE method appeared to be easier for the panelists, however. Essentially, the ISSE approach is the same as the item score procedure proposed by Angoff (Angoff, 1971, p. 514). It requires standard setting panelists to determine the most likely score for a borderline student for each item. The result of the ISSE process is an item score string similar to a string based on student's item responses (Reckase & Bay, 1998).

The use of the IRT-based techniques with the ISSE rating method has these appeals: the numerical achievement standards will be directly on the IRT-based NAEP scale, the method gives differential weights for rating items, and the method accounts for different types of items. The ISSE method and its estimation techniques were also found to be robust with small sample size (Chen & Pommerich, 1998). The ISSE method was field tested for setting ALS achievement levels for the 1998 Civics NAEP and the 1998 Writing NAEP. Analysis outcomes for the field trials, however, raised serious concerns about the bias produced by the ISSE method. The field data suggested that the ISSE method would result in more extreme cutscores than the Mean

Estimation Method, possibly due to the reduced numerical precision of panelist ratings. As a result, the method was not recommended for formal ALS process or further investigations.

The Reckase Method

For item-by-item rating methods to be effective for setting achievement standards, intra-rater consistency plays an important role. Low intra-rater consistency for panelists across rating items indicates poor quality of panelist ratings. It may also indicate that the panelist misunderstood the achievement-levels descriptions or the rating techniques. The Reckase Method (Reckase, 1998a; Reckase, 1998b) was introduced by ACT to provide panelists with useful and easy-to-understand information regarding the consistency in their own ratings in the IRT context, within round and between rounds, and for items of different types, content, or difficulty levels.

The centerpiece of the Reckase method is the Reckase chart, named after the researcher who originally proposed the Reckase method. The Reckase chart is essentially a numerical representation of the item characteristic curves. The entries of numerical values (probabilities or expected scores) in the Reckase chart were generated by the 3-PL IRT model and the Generalized Partial Credit IRT model for dichotomously and polytomously scored items respectively. These numerical entries for individual rating items were arranged in columns side by side, with item-number headings. The discrete ACT NAEP-Like Scale scores, corresponding to the probability or expected score for each of the rating items, were presented in the far left column on the chart in descending order. Using the chart, a panelist could locate his/her item ratings from previous round and find the corresponding ACT NAEP-Like scale score for that item. The pattern of their ratings across items thus became visually clear. The panelists were able to inspect their ratings for each item with respect to their own cutscore and the grade level cutscore, regardless of item type and content.

The 3-PL IRT model used for computing the probability of a correct response for a dichotomously scored item i given ability level θ, as shown in the Reckase Charts, is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (i = 1,2,...,n),$$

where $p_i(\theta)$ is the probability of a correct response for item i given θ,
a_i is the discrimination parameter for item i,
b_i is the difficulty parameter for the item,
c_i is the lower asymptote (guessing) parameter for the item,
and D=1.7 (a scaling factor).

The Generalized Partial Credit IRT model (Muraki, 1992) used for computing the probability of obtaining a given response k for a polytomously-scored constructed response item i is:

$$P_{ik}(\theta) = \frac{e^{\sum_{v=1}^{k} Da_i(\theta - b_i + d_{iv})}}{\sum_{c=1}^{m_i} e^{\sum_{v=1}^{c} Da_i(\theta - b_i + d_{iv})}},$$

where $P_{ik}(\theta)$ is the probability of response k to item i,
m_i is the number of response categories of item i,
a_i is an item discrimination (slope) parameter,
b_i is an item location parameter,
and d_{iv} is a threshold parameter corresponding to the score category v.

Note that in Muraki's Generalized Partial Credit IRT model, the first threshold (d_{i1}) is always constrained to equal zero to solve the indeterminacy problem. The entry for a polytomously scored item on the Reckase Chart is the expected score for that item, given a particular θ value, computed as below:

$$S_i(\theta) = \sum_{k=1}^{m_i} kP_{ik}(\theta).$$

Note that the model shown above has been used for calibrating the NAEP items by ETS.

The Reckase method was field tested for its applicability for the 1998 ALS for the Civics and Writing NAEP (Loomis, et al., 1999). Following the rating process using the modified-Angoff method for dichotomously scored items and the Mean Scores Estimation method for the polytomously scored items, the Reckase charts were presented to panelists. The panelists then marked their own item ratings from previous round on the charts. They also drew lines representing individual and grade-level cutpoints across items on the charts. After the panelists reviewed their own item ratings from previous round with their own cutpoint estimates and the grade-level cutpoint estimates, with the visual aid from the Reckase charts, they were asked to modify their item ratings based on their review.

Although the Reckase method was designed to provide information regarding rating consistency across items within round for the ALS panelists and to help adjust ratings to be more in line with the ICC curves, the 1998 field trial data did not sufficiently demonstrate the positive effect of the Reckase method. The desired improvement in panelists' item ratings due to the method was not obvious. The possible impact of the Reckase method on panelists' item ratings was not clear. It was even concerned that panelist ratings might be driven by the information provided by the Reckase Charts (or the ICC curves). These worries would remain legitimate

until the method was further scrutinized. Therefore, the Reckase method as a rating approach was not recommended for further use for the ALS process. However, ACT decided to provide individually customized Reckase charts (with their ratings and cutpoints from previous round electronically marked and drawn across items) to the standard setting panelists for the pilot studies and the formal 1998 ALS processes (Loomis, 1998a; Loomis, 1998b; Loomis, et al., 1999). For the 1998 Civics NAEP and the 1998 Writing NAEP, the Reckase charts provided additional feedback information that were useful for evaluating the intra-rater consistency for the panelists. The charts also provided valuable information within and across rounds of item ratings for research purposes (Yang, 2000).

Holistic Approaches to Setting Standards

Standard setting based on item-by-item ratings is criticized for being cognitively complex for the judges. In an effort to ease cognitive burden for the standard setting judges to improve the standard setting outcome, holistic approaches to setting standards have been adapted and used for the NAEP ALS process over years. The Paper Selection method, the Booklet Classification method and the Item Mapping method represent the three major efforts of ACT for setting standards from a holistic perspective.

Paper Selection

The paper selection method was developed particularly for setting standards for the polytomously scored items. For each polytomously scored item, the method required panelists to select a number of papers representing student performance at the lower borderline of each achievement level from a pool of student paper samples. The sample of student papers included a variety of student's responses that covered a full range of score points for the same performance-type tasks. The student papers were scored prior to the selection process. However, the panelists reviewed the student papers without knowing their scores. Scoring rubrics were provided to judges during the review process (ACT, 1993). For each rating item, the average score for the selected papers was computed to represent a panelist's estimate of the expected score for a lower borderline student. The estimation of a panelist's cutpoint on the θ scale for a polytomously scored item was based on the Generalized Partial Credit IRT model (Muraki, 1992).

Note that the Paper Selection method yields discrete score points for the computation of cutpoints. The method requires the panelists to select one or up to a small number of papers scored with discrete points. The Mean Scores Estimation method, however, asks a panelist to provide direct estimate of the mean score for a rating item on a continuous scale. For the Paper Selection method, a panelist's ability to apply the scoring rubrics for selecting papers is crucial. Using the Paper Selection method, another source of estimation error for the cutpoint comes from the inaccuracy in the scoring of the papers prior to the selection process. The implication is that the standard error of estimate for the cutpoint is likely to be large (Reckase & Bay, 1998). Despite that the panelists from previous ALS studies indicated that they understood the Paper Selection method and were comfortable with the resulting outcomes of achievement levels, the completion of each round of ratings took a long time. In addition, the number of papers for which the judges needed to review might be excessively large, and the inclusion of the sample

student papers based on a particular score distribution may lead to different outcomes of panelist review. The limited number of papers selected by panelists for estimating panelist achievement standards may introduce another source of error.

The ALS process conducted in 1990 was not used to set the cutscores for the 1990 Mathematics NAEP. In 1992, ACT used the modified Angoff method and the Paper Selection method to set cutscores for the reporting of the 1990 Mathematics NAEP data. The Paper Selection method was not used for setting the NAEP achievement levels in the following years, due to the issues discussed above. Instead, it was adapted for the ALS as part of the training process for the rating panelists.

Booklet Classification

In addition to its application to the field trial for the 1998 Writing NAEP ALS process (Loomis, et al., 1999), the Booklet Classification method was employed by ACT for validating various ALS outcomes (ACT, 1995; ACT, 1997; Hanson, Bay, & Loomis, 1998). The method was further examined, using a different estimation procedure for computing achievement-level cut points, for its applicability for the formal ALS process (Hanson, 1998).

The Booklet Classification method required panelists to classify a sample of student booklets into the three NAEP achievement levels, namely Basic, Proficient, and Advanced, plus the Below-Basic level. The student booklets were selected with considerations of NAEP test content and block types, and the distribution of actual student performance on the NAEP test. Panelist classification outcomes were used with information about student ability (θ) to set the ALS cutpoints. Plausible values derived by ETS for scaling purpose for the NAEP and other reasonable ability estimates were used to arrive at information regarding θ.

To validate the ALS cutpoints for the 1996 NAEP Science, one diagnostic approach and one nonparametric discriminant analysis were developed for estimating individual panelist's cutpoints and grade-level cutpoints (Hanson, Bay, & Loomis, 1998). To refine the standard setting methodology for setting the 1998 NAEP Writing achievement levels, a collapsed-categories method and a borderline-categories method were developed. These two methods were used with the Booklet Classification data from a field trial for estimating cutpoints on the ACT NAEP-like score scale. A weighted combination of the cutpoints resulted form these two methods was used to represent the final cutpoints for the field trial data. In addition, a Cubic Regression approach (Plake & Hambleton, 1998; Hanson, 1998) was applied with the Booklet Classification method to examine data collected from the field trial.

Various study outcomes have indicated that the cutpoints resulting from the Booklet Classification method were likely to be higher than the cutpoints estimated by the Mean Scores Estimation method. The methodology used to derive the numerical cutpoints from the classification data remain a controversial issue. Some other issues concerning the applicability of the Booklet Classification method are: the sufficient number of booklets to be classified by panelists, the number of categories for the classification task, the appropriate score distribution used for selecting booklets for the study, and the method used for estimating the latent scores for the booklets (Loomis, et al., 1999).

Item Mapping

The item mapping procedure is a variation of the Bookmark procedure developed by Lewis, Mitzel, and Green (1996). The method requires items to be ordered through the 3-PL IRT model on a θ scale, and the θ estimates for test items should all correspond to one single probability of correct response (the response criterion). The value for the response criterion is arbitrary but should be reasonable. During the standard setting process using item mapping, a panelist reviewed the ordered list of test items and selected two items on the list that were considered closest to the cutpoint between two achievement levels. The θ values corresponding to these two items were then averaged and the average θ is used as an estimate of the cutpoint for the panelist. For each achievement level, the estimated grade-level cutpoint is the simple average of the estimated individual cutpoints using the item mapping approach. Compared to the modified-Angoff method, the item mapping method may yield larger standard errors for the estimates of the achievement standards. It is because the item mapping method practically relies on information from only two items to estimate the achievement-level cutpoint for a panelist.

The Item Mapping method was not only used previously by ACT for research purposes, but it was also applied to a formal achievement-levels setting process. When the grade 8 panel was reconvened for modifying their ALS achievement levels for the 1996 Science NAEP, the Item Mapping method was used to replace the combination of the modified-Angoff method and the Mean Scores Estimation method. In addition, the Item Mapping method was adapted and tried out in conjunction with the modified-Angoff method and the Mean Scores estimation method for the 1998 NAEP Civics ALS. For the first two rounds of the ALS process, the panelists rated the items using the modified-Angoff method and the Mean Scores Estimation method. For the third round, the panelists reviewed the item maps with other feedback information, including the group cutpoints resulted from the previous round of ratings. They then decided whether they would like to modify the group cutpoints and recorded their recommends for the group cut points on the item maps. The final group cutpoints were computed by taking the average of the panelists recommendations.

Recall that the Item Mapping method arbitrarily chooses one single criterion of response probability to determine the order of items on the θ scale. The reasonableness of the selected response probability for setting standards via the Item Mapping approach is often not well justified. Item maps produced from different response probabilities are likely to order items differently and hence results in different cutpoint estimates for the same panelist. Despite this inherent limitation of the Item Mapping methodology, the method remains an informative technique for empirically classifying rating items into various achievement levels.

From Item Ratings to Cutpoint Estimates

The review of the methodologies implemented by ACT to improve the ALS process illustrated the importance and significance of the IRT application for setting achievement levels for the NAEP. Taking into account the IRT calibration for the NAEP assessment, ACT proposed to train ALS panelists and provide them feedback in the context of IRT. The method of

collecting panelists' judgement data and the techniques for converting the data to achievement-levels cutpoints were also developed in consideration of the IRT application.

A crucial part for standard setting is the mapping of the judgement data on the score scale, where cut points are set. ACT has been using the modified-Angoff method for the NAEP ALS over the past nine years, and the Mean Scores Estimation method also have being used for recent years for the polytomously scored items. These two methods are compliant with the IRT component of the NAEP assessment. Therefore, this paper will focus on these two methods for describing the techniques developed for converting judgement data to numerical cutpoints that represent achievement levels.

With the modified-Angoff method and the Mean Scores Estimation method, each panelist's cutscore on the θ scale for the set of rating items is estimated. Several computational procedures are developed for converting item ratings to θ representing a panelist's achievement-level cutpoint. These computational procedures are summarized and discussed below.

The IRT True Score Approach

An approach implemented by ACT uses test characteristic curve (TCC) to derive a panelist's cutpoint from the panelist ratings across items. The sum of the probability estimates for all of the rating items for panelist j, $\Sigma P(\theta_j)$, is used as an estimate of the achievement standard for that panelist on the set of rating items. The θ value corresponding to the estimated numerical achievement standard for the panelist can be found via the test characteristic function.

The IRT true score approach assumes that panelists' item ratings represent the IRT true scores for the borderline performance for each achievement level. The task then is to find the θ corresponding to the true score Π_j. The θ should satisfy the following equality:

$$\Pi_j = \sum_{i=1}^{n} \pi_{ij} = \sum_{i=1}^{n} \sum_{k=1}^{m_i} kP_i(\theta_j; \xi_i),$$

where π_{ij} is the rating for item i of panelist j, and ξ_i are the item parameters.

For each panelist, the above equation produces an estimate of cutpoint on the θ scale. The overall group-level cutpoint can be a simple or weighted average of the θ estimates across panelists. An information-weighting method for deriving group-level cutpoint will be discussed later.

Alternatively, we can take the average of item ratings across panelists for each item and treat the average as the true score for each item. Summing the true score estimates across items,

we obtain one single estimate for the overall cutpoint. Given a set of n rating items and a number of N panelists, the task is to find θ that satisfies the following equation:

$$\sum_{i=1}^{n} \frac{\sum_{j=1}^{N} \pi_{ij}}{N} = \sum_{i=1}^{n} \sum_{k=1}^{m_i} kP_i(\theta;\xi_i).$$

The above alternative approach is preferred over the former approach because of its smaller estimation error. The average item rating across panelists is an unbiased estimate of the overall rating, and the corresponding overall θ estimate contains only estimation error from the average item rating. For the former approach, however, the resulting overall cutpoint will associate with a larger estimation error than the alternative approach because each of the individual estimates for θ_k contains estimation error for the item parameters ξ.

Despite the advantage described above, the average-rating-based alternative fails to inform individual panelist about his/her cutpoint resulting from his/her item ratings. Therefore, ACT decided to use the alternative average-rating approach for deriving overall cutpoint for each achievement level but use the former approach in estimating individual θ values for panelists. The individual θ estimates were provided to the panelists as feedback between rounds of ratings.

A precision/information weighting approach was proposed for combining the estimates of achievement standards across panelists (Luecht, 1993a). The approach would weight the individually estimated cutpoints for panelists proportionally by their measurement precision. The weighting method was designed to minimize the variance for the composite standard across panelists. The minimized variance would implicitly assure a high degree of consensus across panelists. Since the IRT item information function is related to the standard error of measurement at any ability level, the variance of the cutpoint estimate for a panelist could be approximated by computing the reciprocal of the sum of the item information, conditioned on the judge's true standard estimate. Other than the above approach, Luecht (1993) argued that the information-weighted composite estimate of θ across panelists could be more directly derived by averaging the individual sums of the item ratings over panelists and relating that mean sum of ratings to a point of the test characteristic curve.

The Maximum Likelihood Procedure

A maximum likelihood procedure was developed for achievement levels setting for the NAEP. For this procedure, panelists ratings are transformed to the logit metric to approach a nearly normal distribution for ratings and to achieve equal variances across panelists for the errors in prediction of the observed item probabilities from the estimated group standard (Davey, Fan, & Reckase, 1996). The underlying assumption of the procedure is that each judge's transformed probability ratings on the logit metric form a sample from a normal joint distribution

of ability, with a mean ability value $\mu_i(\theta)$ and variances σ_i^2. Therefore, for each judge, the objective of the maximum likelihood procedure is to find a value for the ability parameter that maximizes the likelihood of the observed ratings. The group standard is estimated by finding the ability value that maximizes the likelihood of the observed ratings, via iterative numerical methods, over items and judges:

$$L(\theta \mid r_{ij}') = \prod_{i=1}^{n} \prod_{j=1}^{N} f_i^{\theta}(r_{ij}'),$$

where θ is the value for the ability parameter,

r_{ij}' is the logit of the probability estimate of judge j on item i ($r_{ij}' = \ln \frac{r_{ij}}{1 - r_{ij}}$),

n is the total number of rating items,

N is the total number of panelists,

and $f_i\theta$ is the joint distribution of ability, with a parameter value θ, for item i.

In estimating the maximum likelihood θ, Davey, Fan, and Reckase (1996) relaxed the assumption of equal variances such that the variance σ_i^2 could vary across items. However, to be practical, they empirically grouped items into four homogeneous subsets and estimated a common variance for each item group. The grouping of items was based on some residual variances for items (see Davey, Fan, & Reckase, 1996).

Although the MLE method was first applied for improving the ALS outcomes in 1996, it was not used for the 1996 ALS process. The Technical Advisory Committee on Standard Setting (TACSS) for the ACT NAEP ALS project did not recommend its use before the ALS process because results for the 1996 ALS process would not be comparable to those from previous years. After the 1996 ALS process, however, NAGB reviewed the cutpoints resulted from both the IRT true score method and the MLE approach and decided to accept the outcomes based on the MLE method.

The Bayesian Procedure

The above maximum likelihood procedure resulted in a single point estimate (the mode) of the achievement level. To adequately characterize the latent distribution of theta, a Bayesian procedure was further developed to provide posterior distributions of achievement standards for each of the ability level. Individual posterior distributions across raters were summed to construct a single, joint posterior distribution (Davey, Fan, & Reckase, 1996).

Both the Bayesian and the Maximum Likelihood procedures have been studied by ACT as an effort in improving the estimation of achievement levels. Subsets of the ALS data for the 1992 NAEP Reading and Mathematics, and the 1994 NAEP Geography, were analyzed using the Maximum Likelihood and the Bayesian procedures respectively to understand the applicability

and efficiency of these procedures. For polytomously scored items, the Maximum Likelihood procedure and the Bayesian procedure require that panelist's item scores be transformed to the 0-1 scale, followed by the logit transformation. The scale is transformed by dividing panelist's item score by the maximum possible score point for each item. Also, the polytomously scored items should be calibrated with a polytomous latent trait model.

Using stability of the estimation procedures as an evaluation criterion, Davey, Fan, & Reckase (1996) found that both the maximum likelihood procedure and the Bayesian procedure resulted in smaller standard errors than the IRT true score-based procedure. In summary, the two procedures were developed to be consistent with the IRT-based assessment of NAEP. By taking into account the variability in the characteristics of the test items, these two procedures attempted to minimize the error in estimating θ. They are more likely to weight less erratic judges and poorly discriminating items, and therefore result in smaller standard errors than the IRT true score methods. Unlike the IRT true score approach, which was implicitly based on the Rasch model, the maximum likelihood and the Bayesian procedures are based on the 3-PL IRT mode and thus are theoretically stronger in estimating the θ.

Combining Ratings for Different Types of Items

When the National Assessment Governing Board set achievement levels for the NAEP Mathematics in 1990 (Hambleton & MLB, 1991), panelist item ratings were unit-weighted. One drawback of unit weighting is that it does not differentiate the cognitive demand for constructed-response item from that for multiple-choice item. Another practice used in earlier years was to weight items by their score points. Under such a weighting system, the resulting cutpoints depended on the total score points of the constructed-response items heavily. When the number of constructed-response items increased over years, the cutpoints resulting from such weighting system depended more heavily on the constructed-response items. It was argued that the score-point weighting method might give constructed-response items more weights than it deserved. There was a concern about the relatively high rate of omits or no response and the relatively little information associated with the constructed-response items, due to their relatively high level of difficulty. The 1992 ALS outcomes revealed that the constructed-response items generally had rather flat item characteristic curves and were not very informative for ability estimation.

After a series of research for weighting items of different types, ACT introduced an information-weighting method (Luecht, 1993a) as an alternative for combining ratings for different types of items. After reviewing cutpoints computed by the two weighting methods, NAGB favored the information weighting and accepted its resulting cutpoints. The information-weighting method remained in use until 1996. One major advantage of information weighting is that it takes into account item information, a concept closely related to standard error. Unlike the score-point weighting, constructed-response items that yield little information/discrimination will not count much toward determining the cutscore when information-weighting method is used.

The information-weighting method for combining the composite estimates of achievement-level cutpoints for different types of items is analogous to the precision/information

weighting approach for combining the cutpoint estimates across panelists (discussed earlier in this paper). The required assumption for the weighting method is that all the dichotomously scored item and the polytomously scored items are calibrated or linked to a common metric of latent ability (Luecht, 1993a). A combined composite estimate ($\hat{\theta}_c$) of the achievement standard on the latent metric can then be obtained by weighting each of the component estimates ($\hat{\theta}^*$) of the achievement standard by the amount of information:

$$\hat{\theta}_c = \frac{\sum_{h=1}^{s}\left[\hat{\theta}_h^* \sum_{i=1}^{n_h} I_i(\hat{\theta}_h^*)\right]}{\sum_{h=1}^{s}\left[\sum_{i=1}^{n_h} I_i(\hat{\theta}_h^*)\right]},$$

where s is the number of component estimates.

Note that s=2 for the NAEP ALS practice, where component cutpoints were estimated respectively for the dichotomously scored items and polytomously scored items. Based on the 3-PL IRT model, the information function for the dichotomously scored item i is:

$$I_i(\theta) = \frac{D^2 a_i^{\,2}(1-c_i)^2\left(P_i^* Q_i^*\right)^2}{P_i Q_i},$$

where D=1.7, the adjustment for a normal metric,
P_i is the 3-PL IRT model shown earlier in this paper,
Q_i=1-P_i,

$$P_i^* = \frac{1}{1+e^{-Da_i(\theta-b_i)}},$$

and Q_i^*=1-P_i^*.

Using the Generalized Partial Credit Model, the information function for the polytomously scored item i is:

$$I_i(\theta) = D^2 a_i^{\,2}\left[\sum_{k=1}^{mi} k^2 P_{ik}(\theta) - \left(\sum_{k=1}^{mi} k P_{ik}(\theta)\right)^2\right],$$

where P_{ik} is the Generalized Partial Credit model shown earlier in this paper.

Luecht (1993a) further indicated that the combined composite estimate of the standard across judges would have approximately minimum asymptotic variance properties. Assuming that the covariance between the latent standards for different types of items is zero, the large sample variance of $\hat{\theta}_c$ is approximated by calculating:

$$VAR(\hat{\theta}_c \mid \theta_c) = \frac{1}{\sum_{h=1}^{s}\sum_{i=1}^{n_h} I_i(\hat{\theta}_h^*)}.$$

References

ACT (Formally American College Testing) (1993). Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity. Iowa City, IA: Author.

ACT (1994). Design document: Setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science. Iowa City, IA: Author.

ACT (1995). Research studies on the achievement levels set for the 1994 NAEP in Geography and U.S. History. (Unpublished)

ACT (1997). Achievement levels-setting study. In Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science Final Report. Iowa City, IA: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement (2nd Edition). Washington, DC: American Council on Education.

Chen, W. H., & Pommerich, M. (1998). Setting achievement level standards for NAEP using item score judgment: A simulation study. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Davey, T., Fan, M., & Reckase, M. D. (1996). Some new methods for mapping ratings to the NAEP theta scale to support estimation of NAEP achievement level boundaries. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Hambleton, R. K., & Bourque, M. L. (1991). The levels of mathematics achievement. Vol. III, technical report. Washington, DC: National Assessment Governing Board.

Hanson, B. A. (1998). Application of a cubic regression method for setting NAEP Writing Achievement Levels using booklet classification data. Paper prepared for the Meeting of the Technical Advisory Committee on Standard Setting for the ACT NAEP Project , Iowa City, IA: ACT.

Hanson, B. A., Bay, L., & Loomis, S. C. (1998). Booklet Classification Study. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) Educational Measurement (3rd Edition). New York: American Council on Education & Macmillan.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. Paper presented at the CCSSO National Conference on Large Scale Assessment, Phoenix.

Loomis, S. C. (1998a). NAEP 1998 Civics ALS pilot study overview. Report to the Technical Advisory Committee on Standard Setting, September 1998, Minneapolis.

Loomis, S. C. (1998b). NAEP 1998 Writing ALS pilot study summary. Report to the Technical Advisory Committee on Standard Setting, October 1998, Detroit.

Loomis, S. C., Bay, L., Yang, W., & Hanick, P. L. (1999). Field trials to determine which rating method(s) to use in the 1998 NAEP Achievement Levels-Setting Process for Civics and Writing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Luecht, R. M. (1993a). Using IRT to improve the standard setting process for dichotomous and polytomous items. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Luecht, R. M. (1993b). Some results on the stability of the NAEP Achievement level standards in mathematics. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159-176.

National Assessment Governing Board (1990). Setting appropriate achievement levels for the National Assessment of Educational Progress. Washington, DC: Author.

Plake, B. S., & Hambleton, R. K. (1998). A standard setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Reckase, M. D. (1998a). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. Applied Measurement in Education, 11 (1), 9-21.

Reckase, M.D. (1998b). Setting standards to be consistent with an IRT item calibration. Iowa City, IA: ACT.

Reckase, M. D., & Bay, L. (1998). Analysis of methods for collecting test-based judgements. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Sanathanan, L., & Blumenthal, S. (1978). The logistic model and estimation of latent structure. Journal of the American Statistical Association, 73, 794-799.

Yang, W. L. (2000). Analysis of Item Ratings for Ensuring the Procedural Validity of the 1998 NAEP Achievement-Levels Setting. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.